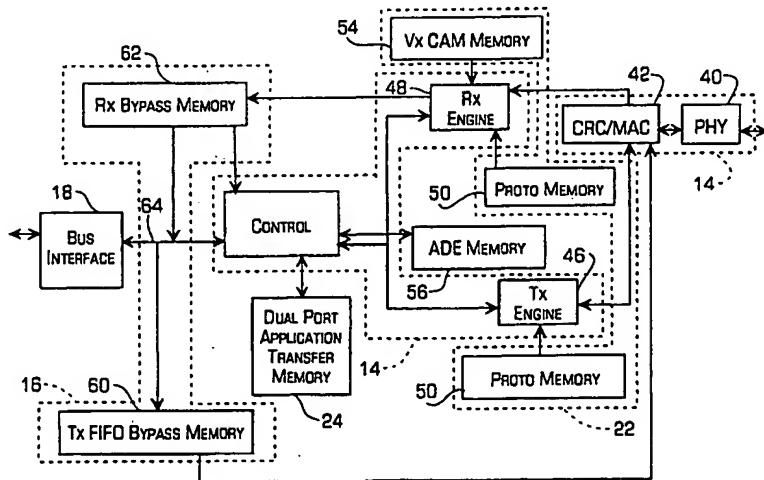




INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

| | | | |
|--|--|----|--|
| (51) International Patent Classification 6 : G06F 13/00 | | A1 | (11) International Publication Number: WO 99/04343 |
| | | | (43) International Publication Date: 28 January 1999 (28.01.99) |
| (21) International Application Number: PCT/US98/14729 | | | |
| (22) International Filing Date: 17 July 1998 (17.07.98) | | | |
| (30) Priority Data: 60/053,240 18 July 1997 (18.07.97) US | | | |
| (71) Applicant (for all designated States except US): INTER-PROPHET CORPORATION [US/US]; Suite 1020, 1999 South Bascom Avenue, Campbell, CA 95008 (US). | | | (81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG). |
| (72) Inventors; and | | | Published |
| (75) Inventors/Applicants (for US only): JOLITZ, Lynne, Greer [US/US]; 22570 Citation Drive, Los Gatos, CA 95033 (US). LAWSON, Matthew, Todd [US/US]; 570 El Camino Real #150-380, Redwood City, CA 94063-1200 (US). JOLITZ, William, Frederick [US/US]; 22570 Citation Drive, Los Gatos, CA 95033 (US). | | | With international search report. |
| (74) Agent: YOUNG, Barry, N.; Gray Cary Ware & Freidenrich, 400 Hamilton Avenue, Palo Alto, CA 94301 (US). | | | |

(54) Title: TCP/IP NETWORK ACCELERATOR SYSTEM AND METHOD



(57) Abstract

A network accelerator (10) for TCP/IP includes programmable logic (46, 48) for performing network protocol processing at network signaling rates. The programmable logic is configured in a parallel pipelined architecture controlled by state machines (100, 140) and implements processing for predictable patterns of the majority of transmissions. Incoming packets are compared with patterns corresponding to classes of transmissions which are stored in a content addressable memory (22), and are simultaneously stored in a dual port, dual bank application memory (24). The patterns are used to determine sessions to which an incoming IP datagram belongs, and data packets stored in the application memory are processed by the programmable logic. Processing of packet headers is performed in parallel and during memory transfers without the necessity of conventional store and forward techniques resulting in a substantial reduction in latency. Packets which constitute exceptions or which have checksum (62) or other errors are processed in software.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

| | | | | | | | |
|----|--------------------------|----|---------------------------------------|----|---|----|--------------------------|
| AL | Albania | ES | Spain | LS | Lesotho | SI | Slovenia |
| AM | Armenia | FI | Finland | LT | Lithuania | SK | Slovakia |
| AT | Austria | FR | France | LU | Luxembourg | SN | Senegal |
| AU | Australia | GA | Gabon | LV | Latvia | SZ | Swaziland |
| AZ | Azerbaijan | GB | United Kingdom | MC | Monaco | TD | Chad |
| BA | Bosnia and Herzegovina | GE | Georgia | MD | Republic of Moldova | TG | Togo |
| BB | Barbados | GH | Ghana | MG | Madagascar | TJ | Tajikistan |
| BE | Belgium | GN | Guinea | MK | The former Yugoslav Republic of Macedonia | TM | Turkmenistan |
| BF | Burkina Faso | GR | Greece | ML | Mali | TR | Turkey |
| BG | Bulgaria | HU | Hungary | MN | Mongolia | TT | Trinidad and Tobago |
| BJ | Benin | IE | Ireland | MR | Mauritania | UA | Ukraine |
| BR | Brazil | IL | Israel | MW | Malawi | UG | Uganda |
| BY | Belarus | IS | Iceland | MX | Mexico | US | United States of America |
| CA | Canada | IT | Italy | NE | Niger | UZ | Uzbekistan |
| CF | Central African Republic | JP | Japan | NL | Netherlands | VN | Viet Nam |
| CG | Congo | KE | Kenya | NO | Norway | YU | Yugoslavia |
| CH | Switzerland | KG | Kyrgyzstan | NZ | New Zealand | ZW | Zimbabwe |
| CI | Côte d'Ivoire | KP | Democratic People's Republic of Korea | PL | Poland | | |
| CM | Cameroon | KR | Republic of Korea | PT | Portugal | | |
| CN | China | KZ | Kazakhstan | RO | Romania | | |
| CU | Cuba | LC | Saint Lucia | RU | Russian Federation | | |
| CZ | Czech Republic | LI | Liechtenstein | SD | Sudan | | |
| DE | Germany | LK | Sri Lanka | SE | Sweden | | |
| DK | Denmark | LR | Liberia | SG | Singapore | | |
| EE | Estonia | | | | | | |

TCP/IP NETWORK ACCELERATOR SYSTEM AND METHOD

Background of the Invention

5

The present invention relates to Internet communications in general, and to a method and system in particular for substantially increasing the data throughput of TCP/IP protocol based data transmissions by selectively implementing in hardware certain portions of the TCP/IP protocol set (such as a majority of actually called and executed routines), and 10 implementing in software routines the exceptions and remaining portions.

10

Since the implementation of FDDI fiber network links, the transmission speed of the physical layer to transmit data, has exceeded the ability of the end node computers to process the data packets. If the processing of the data packets is done by Von Neuman architectured end node computers, capacity is always exceeded since the switching speed of the fastest 15 computer's gates will be approximately equal to that of the physical layer comprising the internal components of Application Specific Integrated Circuit (ASIC) chips. The computer CPU (which must process the data packets with multiple operations and copies to memory) intrinsically requires orders of magnitude more device operations than that of the analog/state machine mediated physical layer of the ASIC chips normalized to a common amount of data.

20

While the problem of scaling current computer networks to gigabit speeds has been recognized, the complexity of the TCP/IP protocols has presented both practical and conceptual barriers to attempts to implement them in any manner other than various forms of software executed processes. However, even the fastest of CPUs for any given technological generation, cannot match the physical bandwidth of their internal components.

25

There have been a number of attempts to accelerate TCP/IP protocol handling, but none has effectively solved the latency problems. One approach to accelerate TCP/IP protocol handling was to process the headers of the protocols independently of the data payload. While the implementation of the protocols themselves was virtually identical to existing methods (TCP/IP software stack), the data was indirectly manipulated by separate

buffering to avoid multiple copies of the payload data through the use of hardware buffer management using a multi-port memory. This approach demonstrated that hardware buffer management could improve handling of large payload packets, but it did not reduce packet latency to memory, did not improve the control bandwidth of the protocol or the ability to 5 send small packets efficiently, and did not decouple protocol processing speed from transmission speed. The approach also was not applicable to local clusters, or to small record applications like web-serving or transaction processing. Moreover, the approach did not eliminate the store/forward processing of protocols, but merely attempted to optimize the methods by which the store and forward were mediated.

10 ATM cell-based transmission technology incurs a cost because of segmentation and reassembly of large data payload messages into much smaller cells. Devices which attempt to minimize this cost perform this function at the signaling rate. However, this function is specific to cell-based technologies, and is not particularly useful for technologies such as Ethernet and HiPPI. The payload size of such technologies' packets do not require an 15 adaptation layer below that of the network or IP (Internet Protocol) layer. In order to process TCP/IP protocols, traditional store and forward methods must be used.

Protocol engines have also been used to optimize traditional methods of protocol handling to reduce certain steps. These include hardware checksum units, hardware buffer management, and RISC processing to improve protocol handling rate. However, this 20 approach still does not scale with signaling rate.

Other approaches have implemented in hardware proprietary non-TCP/IP protocols having a continuous flow and routing that is specific to the particular network fabric. Variable context matching is not performed, and cells propagate in strict format and order to a priori known memory addresses instead of to a transport protocol's abstract port destination. 25 Therefore, such approaches are not readily adaptable to wide area networks which must handle a variable and relatively unstructured traffic flow, and which must be scaleable, expandable and readily adaptable to network changes.

It is desirable to provide a network accelerator system and method for handling standard TCP/IP protocol which solves the latency and other problems of known systems and 30 methods, and it is to these ends that the present invention is directed.

Summary of the Invention

The present invention provides a solution to the above-mentioned protocol processing problems using a cross disciplinary combination of hardware elements, techniques and results based, inter alia, on network traffic analysis, high speed programmable logic array technology, and integration with low level operating system software design.

The invention solves a problem that has been long unsolved of how to process TCP/IP data packets at a speed equal to that made possible by the latest generation physical layer hardware transmission components. As microprocessors increase in speed, the same technology advances also increase the speed at which data can be transmitted over networks. If this data protocol handling must be handled in software, then there are fundamental issues in logic and software design that will always make the ability of a processor to process the packets slower than the physical ability of the network to transmit packets. This speed differential can penalize maximum possible network performance by a factor of almost one hundred at present.

The main insights that enable the invention to provide a practical and implementable solution to the above-mentioned protocol processing problems are the recognition that the transmission patterns of the vast majority of packets over current TCP/IP mediated networks are predictable and involve only a very small subset of the entire TCP/IP protocol set. It is possible through logic design to implement this small set of actually used protocols in hardware, such as programmable logic gate arrays, to allow processing of TCP/IP data packets at speeds equal to that of the ability of the fastest physical network layer. The rare packets that cannot be handled in this manner can be defaulted to conventional software processing. An operating system also can be low-level interfaced to this processing system through appropriate memory management in such a way that the packet's data coming off the network data transmission medium can be processed and put into application memory at the speed equivalent to a single gate-mediated operation.

The invention allows practical processing of TCP/IP data packets in gate array hardware at a data throughput equal to that of the physical transmission media. It

5 accomplishes this task by recognizing that TCP/IP packets on current networks fall into predictable transmission patterns that actually utilize only a small fraction of the entire protocol for the vast majority of transmissions. By implementing this small subset in gate array hardware and defaulting the exceptions into software, a very large increase in TCP/IP packet throughput can be obtained.

10 TCP/IP transmissions handled by the invention can be made faster than that possible with the best current software implementations and multiprocessor TCP/IP processing engines. Using mask programmable logic affords approaches which are both faster and less expensive to construct than the current RISC CPU assisted TCP/IP processing boards, the invention is intrinsically scaleable upwards in speed with little or no redesign needed as advances in IC processing technology makes the network physical layers faster. A form of software embedded in hardware which can be physically implemented at any point where TCP/IP packet processing is used such as in network interface cards, and within microprocessor CPUs, affording significant potential technological and economic benefits.

15 A difference between the invention and prior approaches is that the invention constructs a path into memory for a specific class of packets that exists for the likely time interval when such a packet will be present. The path into and out of memory is handled entirely in the hardware of the invention with only random logic up to where it interacts with the application, and is triggered entirely by the arrival of the packet itself. In this hardware, 20 all details are present for handling the packet payload state to where it will be delivered. With accelerators on both ends of a network transfer, no software overhead need be present for bulk data transfer in burst mode. This differs markedly from prior software and hardware approaches which employed techniques of minimized protocol implementations, buffer management, or by spreading the protocol implementation across a specially designed 25 network fabric.

30 The invention implements continuous flow (streamed) information delivery via a standard protocol such as (TCP/IP) by means of a pattern match via associative memory. It has several benefits in processing standard protocols, as opposed to non-standard protocols. These include absolute minimum latency between application and network medium (fibre), absolute maximum bandwidth between communicating network applications, low complexity

-5 -

design network protocol processing mechanism, and the protocol rate scales linearly with network signaling rate.

These and other benefits are obtained, in one aspect, by avoiding software and hardware processing steps via an isochronous "stimulus/response" architecture using a variable content addressable memory that has preprogrammed state logic that effects protocol processing as a minimum time series of operations. A substantial, e.g., ten-fold, improvement in interapplication bandwidth with same complexity hardware results which makes practical low-cost gigabit network transport communications. While standard protocol processing is not unique as a process, this inventive method of processing is unique in that the software of a protocol implementation processes protocol information indirectly via hardware which has been a priori instructed on how to handle a predicted flow of packets autonomously. This methodology is superior to prior attempts in that the transmission speed of the network transport layer is scaled with the network physical layer.

15 **Brief Description of the Drawings**

Figure 1 is a diagram of a network accelerator in accordance with the invention;

20 Figures 2a and 2b are diagrammatic views which contrast, respectively, a traditional link data stream store and forward approach with a network accelerated continuous flow link data stream approach in accordance with the invention;

Figure 3 is a more detailed diagram of the network accelerator of Figure 1;

Figure 4 is a diagram of a control unit of the network accelerator of Figure 3;

25

Figure 5 is a diagram of a transmit engine of the network accelerator; and

Figure 6 is a diagram of a receive engine of the network accelerator.

Description of Preferred Embodiments

The invention is particularly adaptable to TCP/IP protocols and will be described in that context. It will be appreciated, however, that the invention has greater utility and is applicable to other streamed protocols.

5 Computer networks use network software protocols to communicate information reliably between computers over multiple successive physical signaling mediums. These protocols are implemented in software on computer processors. While hardware signaling rates have steadily increased, software protocol processing has not kept pace. With the advent of gigabit networking technology, costly processors must be dedicated to providing at 10 most 40-50 percent of the theoretical bandwidth of the network, while software implementations used with earlier signaling technologies were capable of 60-80 percent of theoretical bandwidth. Clearly bandwidth demands will continue to increase, and since the disparity between software protocol processing and signaling rates will also increase, this defines a "bottleneck" in the effectiveness of networking technology.

15 The invention affords a minimum time mechanism for handling TCP "burst" transfers. A burst transfer is a series of bulk data transfers with no options between nodes (usually it is unidirectional). It consists of the sending node passing data payload packets with successive sequence numbers, and the receiver committing them to memory and sending acknowledgments back to the sender to trigger more data to be sent. The invention efficiently 20 handles the burst so as to minimize latency. The software triggers the burst mode by a traditional send operation with a full-payload, and the receiver/transmitter fall into an asynchronous feedback loop of "send-next"/"acknowledgment" packets that continues until either the burst completes or an error occurs.

25 The invention provides a mechanism to process the costly portions of standard protocols in hardware entirely, and to do so at the same clock rate of the signaling. In this way, as the signaling rate rises, so does the protocol processing rate increase in lockstep. This approach is based upon several observations, including traffic pattern analysis of packets, experience with software protocol implementations, and experience with other non-standard hardware-implemented protocols.

Traffic observation of TCP/IP packets shows that the majority of the packets simply pass bulk data without event, while the minority packets require more elaborate handling. Even more significant is that the delays on the Internet are for these very bulk data packets, so that it is critical to have timely delivery or low-latency of these packets for the performance to be maximized. Loss of this low-latency also impacts the reliability of a network, since it becomes impossible to tell if a failure has occurred, or if an assemblage of worst-case delays has masked an otherwise successful transfer. The ability to handle protocol packets with deterministic response time (as well as with a broad range of arrival distributions) is a requirement to maintain the "real time" characteristics that telecommunications services like telephone systems use to provide high-valued services en masse globally.

Experiences with software protocol implementations have shown that the necessary operations for a TCP/IP "burst" mode are constrained enough to be performed by hardware as clocked by the data stream. Unfortunately the difficulty in synchronizing the software with the data stream renders this observation useless. However, significant performance advantage can be gained by relying on hardware logic gate delays instead of program instructions for substantially reducing the latency between the network and application, thus allowing protocol handling at sustained rates without the need for additional buffering. This allows for continuous protocol processing at the data rate of the signaling technology. As will be described, the network accelerator of the invention uses a deterministic state machine to implement the transport protocol bulk receive and transmit functions, leaving to the software all other features of the protocol (including error recovery).

Figure 1 is a functional diagram of a network protocol accelerator 10 in accordance with the invention. As will be described, the network accelerator can perform 100baseTx full-duplex interface to an Ethernet network, with media access controller (MAC functions), IP (Internet Protocol) processing and decoding, and TCP (Transmission Control Protocol) processing. It may be a PCI interface board, designed to be used in an NT workstation, for example, utilizing a standard PCI bus slot. The accelerator will preferably have a physical link layer processor, an IP processor, a TCP processor, segment buffer memory, multiple FPGAs for logic, and a PCI interface to the host system.

As shown in the figure, the network accelerator includes a network interface 12 that includes a physical (PHY) media framing unit which obtains physical signals from the physical media, decodes the signals and the link layer framing as a byte stream, and supplies the stream to an accelerator engine 14. Simultaneously, a copy of the signals may be

5 recorded in a receive/transmit (Rx/Tx) FIFO bypass unit 16. In the event of a failure of the accelerator engine 14 to accept the packet, a system bus interface unit 18 may signal a system interface unit 20 to handle the packet which is stored in the Rx FIFO bypass portion of 16. Similarly, the system interface can send packets via the bus interface to the Rx/Tx FIFO bypass portion of unit 16, which hands it on to the physical media framing unit 12, effectively

10 bypassing the accelerator engine for non-TCP data transfers. The accelerator engine 14 is connected to a variable content addressable memory 22, and consults the memory as octets of a packet are received to find a match with predetermined patterns. When a match is found, a state machine associated with the pattern is loaded from the content addressable memory into the accelerator engine to operate on the packet. Operations may include packet table delivery

15 from the physical media framing unit to a dual port application transfer buffer 24; sending a packet with payload from the application transfer buffer to the physical media framing unit; and sending a packet acknowledgment to the physical media framing unit.

Upon completion of an operation, the accelerator engine 14 may signal its status to the system interface unit 20 via the bus interface 18 indicating the event. In the event the

20 accelerator engine fails to recognize a packet or encounters an error, such as a CRC (cyclic redundancy check) or checksum error, the accelerator engine suspends operation on that segment causing all packet traffic to be handled via the Rx/Tx FIFO bypass buffer until it is re-enabled by the system interface unit 20 via the bus interface 18 to return to normal operation.

25 Preferably, the network accelerator of the invention is implemented as a set of programmable logic chips intermediate between the network physical layer interface chip set and an interface chip set for the PCI or other bus. These programmable logic chips may be SRAM field programmable gate arrays (FPGAs) which can be reprogrammed via the network to allow the hardware protocols to be modified after installation to correct errors, to optimize

30 performance as the network changes, or to implement changes to existing protocol sets. Low

cost implementations can also use mask programmable chip sets. The speed advantages of mask programmable ASICs may make their use preferable in high speed point to point data transfer applications where the end nodes and routers are well defined by the user.

5 The network accelerator also may be implemented within the silicon of the main microprocessor or within on-board multi-chip-modules analogous to the way MMX incorporates digital signal processor functionality, or by which AGP provides on-chip integration of graphics accelerator functions. This can bring the data directly into the microprocessor, bypassing the external data bus interface which otherwise limits performance.

10 The high speed data transmission capability of the invention is advantageous in providing for direct data storage, display, or data processing device interconnect both between and within individual computers.

15 Figures 2a and 2b are diagrammatic views which contrast the significant improvement in latency between a network accelerated continuous flow link data stream approach of the invention (Figure 2b) and a traditional link data stream store and forward approach (Figure 2a). Figure 2 illustrates how a traditional protocol stack accumulates data in a store and forward buffer 30, and then performs the necessary protocol processing operations. As indicated in the figure, data packets from an Ethernet are delivered to a link data delivery unit which may perform error checking prior to storing in buffer 30. The time required for this 20 operation is of the order of tens of microseconds. Subsequently, the various segments of the data packet are processed in a protocol processor 32. This processing is sequential, and results may also be placed in the store and forward buffer as application payload data is delivered to the protocol processor. This typically may take hundreds of microseconds, even with very high speed devices performing the operations. The critical weakness is that data 25 must be in some kind of buffer before it can be processed, and processing must be completed before the data can be forwarded.

30 In contrast, as shown in Figure 2b, the network accelerator of the invention uses the protocol's data stream itself as a way of instructing a uniquely constructed data flow processing machine 34 that is clocked by the protocol data and which performs processing operations as the information appears. As indicated in the figure, processing occurs in a

series of parallel functional units 35-38 having a pipelined architecture so that packets are processed in real time with the processed data flowing between the network's wire link and the processing application's data origin. In effect, a protocol's packet would appear as a single, fat instruction that would run on a data flow processor in lock step with the link's data rate. This allows complete processing in times of the order of tens of microseconds in contrast to the traditional store and forward approach illustrated in Figure 2a. The manner in which this is accomplished will be described in more detail below.

Figure 3 is a functional diagram which illustrates in more detail a preferred embodiment of the network accelerator of Figure 1. As shown, the physical media framing network interface 12 may comprise a physical device interface (PHY) 40 connected to a CRC/MAC unit 42. This physical device interface and CRC/MAC unit provide physical and link layer access, respectively, to an Ethernet network. The CRC/MAC unit provides parallel-to-serial conversion, CRC (cyclic redundancy check) generation and checking, MAC address recognition, FIFO buffering, and interface to the remainder of the network accelerator, which includes TCP/IP processors and the dual port application transfer memory 24, which preferably comprises a dual port/double banked RAM.

As will be described more fully, outgoing Ethernet packets will be read from the buffer memory 24 and transferred to an internal FIFO in preparation for transmission to the network. As the Ethernet packet is constructed and output to the network from the FIFO by the Tx engine, the CRC will be calculated on the fly and appended to the end of the Ethernet packet. An incoming Ethernet packet is stored in the incoming FIFO while the destination address is checked against the MAC address register. If the MAC address is correct, the Ethernet packet is sent to an Rx engine. The Ethernet packet is also run through the CRC checker, simultaneously. Once the Ethernet packet is completely received and the CRC is good, the CRC good signal will be asserted.

As shown in Figure 3, the accelerator engine 14 includes a control unit 44, a Tx engine 46, and a Rx engine 48. Included within the variable content addressable memory 22 is a first prototype memory 50 connected to the Tx engine 46, and a second prototype memory 52 connected to the Rx engine 48. In addition, variable content addressable memory (VxCAM) 22 includes a content addressable memory (CAM) 54 which is also

connected to the Rx engine 48. The variable content addressable memory matches a variety of packet formats and is used to quickly determine to which session an incoming IP datagram belongs. The variable content addressable memory 22 also includes an ADE memory 56 which is connected to control unit 44. The Rx/Tx FIFO bypass memory 16 may be 5 implemented as a Tx bypass memory 60 and a Rx bypass memory 62. As shown, bypass memory 62 may be connected to the Rx engine 48 and to a bus 64 connecting the control unit 44 and bus interface unit 18. The Tx bypass memory 60 may be similarly connected to bus 64 and to the CRC/MAC unit 42.

10 The network accelerator handles the various layers of an Ethernet packet as it is sent or received from a network. When processing the IP layer, the IP address, IP checksums, ID field, flags, IP datagram length, etc. are either pre-calculated and sent to the network via the Tx engine 46, or used to verify the destination of an incoming Ethernet packet via the Rx engine 48.

15 When processing the TCP layer, TCP ports, TCP checksums, sequence numbers, ACK number, flags, window size, urgent pointer, options, etc. are either pre-calculated and sent to the network via the Tx engine, or used to verify the destination of a incoming IP datagram via the Rx engine. The Tx engine 46 obtains the TCP payload directly from the memory 24. The Rx engine 48 delivers the TCP payload directly to the memory 24.

20 The memory 24 will contain the host system view of network memory, and a shadowed copy for the network accelerator to use for TCP segment transmission and reception. The host system software driver will swap application memory (system RAM) for memory 24. This will allow the host system direct access the network data stored in the dual-port / double banked memory, effectively replacing the role of host system RAM. Finally, the system interface controls the relationship between the system and the network accelerator. 25 It contains configuration and status registers, and allows the host system to access the network accelerator.

30 Data for the packets are buffered for transfer using the memory 24. This memory maintains an up-to-date copy of the network data for the host system / application, and a local copy of the network data for the network accelerator. This allows the application/ host system to access memory as it would system RAM before, during and after a TCP segment is

sent to the network by the network accelerator. Also, the memory allows access to a stable copy of the network data for transmission or reception to/from the network. The network acceleration control unit maintains the proper relationship between the memory banks, with the banks synchronized in the case of Idle state (the network accelerator is neither 5 transmitting nor receiving TCP segments), or logically separated during network accelerator TCP segment transmission or reception. The double banked nature of the memory allows a "zero-copy" or "zero-latency" method of network data delivery to the network accelerator.

Along with the bulk memory, there is status memory used to maintain the relationship 10 between the memory bank and the host system memory bank. This status memory works as a table indicating which bank of memory has the most current byte of network data for each address in the memory.

The content addressable memory (CAM) 54 is used to quickly determine to which session an incoming IP datagram belongs. It cooperates with ADE memory 56, prototype memories 50, 52, and is part of the variable content addressable memory (VxCAM) 22.

15 Within the ADE memory 56, there will be one or more address descriptor entries (ADEs) which describe the segment details such as memory base address, TCP payload length, TCP payload checksum, next TCP sequence number and the next TCP segment's ADE. This information is used by the Tx engine when the segment is constructed, prior to transmission. The Rx engine uses the ADE fields to determine the sequence numbers, 20 payload destination, and out-of-order segments.

Within the prototype memories 50, 52, there will be one or more session prototype 25 description entries. These entries describe the session fields that do not change, as well as the initial values for the session, such as IP address, TCP ports, protocol fields, base sequence number, first ADE, etc. The Tx engine uses this information to generate the static fields within a session for an outgoing TCP segment. The Rx engine uses this information to determine what TCP session an incoming TCP segment is destined for, and to verify the validity of specific fields in the TCP/IP header.

The content addressable memory 22 stores the address of the potential TCP session 30 prototype entry that describes the session to which an incoming segment belongs. Certain fields in the TCP/IP header are hashed to obtain a value which is used as an address to "look

up" which prototype describes this segment. The memory stores at the "hashed" address another address which points to the prototype data in the prototype memory. If the memory returns a value of zero, the incoming TCP segment does not belong to any accelerated sessions and is routed to the bypass FIFO. In this manner, a "one shot" lookup of the TCP 5 session prototype can be done, rather than searching potentially thousands of TCP session prototypes.

Figure 4 illustrates the network accelerator control unit 44 in more detail. As indicated above, the control unit provides the overall state machines and control registers which control the network accelerator. Logic for controlling the dual port application transfer 10 memory 24 and the Tx and Rx session state machines (to be described) for the Tx engine and the Rx engine, respectively, may be contained in a dual port memory controller 61. Logic for generating a checksum may be contained in a checksum unit 62 which interfaces with ADE memory 56 via an address bus 57 and a data bus 58. After initialization of a current 15 checksum, the ADE memory 56 may be created and used for bounds checking on the host address to obtain the checksum for the desired payload of a TCP segment. This checksum may be loaded into the checksum unit 62. The current value may be stored in memory 24, and a new calculated value may be added to the checksum. The checksum may then be saved either through a write back to the ADE memory 56 and the dual port application transfer 20 memory 24, or if multiple locations require modifications, by iteration. The calculated checksum is then ready for use as the data checksum of the TCP segment.

As shown in Figure 4, a first FIFO buffer 70 may interface the dual port application transfer memory 24 to Tx data from the Tx engine state machine, and a second FIFO buffer 72 may interface memory 24 to Rx data from the Rx engine state machine. Logic for controlling the FIFOs may be contained in the FIFO buffers themselves and used to minimize 25 bus arbitration read/write by an arbitration unit 74. In addition, logic for controlling the Rx engine and the Tx engine, as well as access to their status and control registers, may be contained in a registers and configuration unit 76 which is interfaced to memory 24 and memory controller 61 by an application data bus 77 and an application address bus 78. Arbitration unit 74 also may include logic to control memory access arbitration between the 30 host system and the network accelerator. The network accelerator control unit also maintains

global control of the state machines for each session.

The Tx state machine and the Rx state machine may have the following states. Tx idle is the state prior to sending a Tx buffer to the network. This is the default state and is set up by the software driver. The software driver will also generate the necessary values for the 5 VxCAM for a given buffer space. The host system fills the memory until the session is ready to be transmitted. At this point, the state machine transitions to the Tx pending state.

During the Tx pending state, the dual port memory controller 61 maintains two copies of the data: one for the host system, and one for the Tx engine. Proper data relationship between the host system memory and the network accelerator memory must be maintained to 10 prevent old data from overwriting new host system data, and new host system data from overwriting the data in use by the Tx engine.

In the Tx complete state, if the transmission fails the state machine goes to the Tx re-transmit. If the Tx was a success, the network accelerator will set a success bit and go to the Tx idle state. The network accelerator is now waiting to send out the next segment. In either 15 case, the network accelerator control unit must continue to maintain the proper relationship between the two copies of data.

If the Tx transaction fails, in the Tx re-transmit state, the network accelerator may either attempt to re-transmit the segment, or move to the next session queued for transmission and attempt this segment later.

20 The Rx idle state is the initial state. In this state, the two copies of network data are reconciled. Depending on the outcome of the previous received segments, the host system reads data from either the shadow bank of dual port application transfer memory or the application bank of same memory. If a packet was successfully received, the net payload data stored in the shadow bank of the dual port application transfer memory must be presented to 25 the host system. This is performed on a byte by byte level.

In the Rx pending state, the Rx engine is receiving one or more segments is the current session. Receive data is placed in the proper bank of the memory by the network accelerator control unit.

30 In the Rx complete state, there may be two different scenarios: Rx success or Rx time-out. In the case of success, the success bit for the ADE is set, then the Rx idle state is

entered. In the case of failure, the state machine goes to idle and no changes occur to the memory.

Checksums for the payload of the TCP packet are calculated by the checksum logic 62 as follows. Upon initial setup of the session, the section of memory used by the session is 5 cleared to all zeros. This allows the initial checksum to be initialized to zero for each segment. ADEs are setup for each segment within the session; ADEs contain the starting address, ending address, and checksum for each segment of the session. There may be one or more segments in any session:

During host system writes, the host system presents an address to be accessed. Bounds 10 checking is performed on this address to determine which ADE contains the checksum for this address. The checksum is loaded into the checksum logic and the current (old) value in the memory is subtracted from the checksum. Next, the new data value is added to the checksum.

Upon saving a new checksum, if it is a single location write, the new checksum is 15 written back into the ADE and the new data value is written into the memory. If multiple locations are to be modified, the checksum stays in the checksum generator and each new data value is added to the checksum while each old value is subtracted. The new data is written into the dual port application transfer memory 24 during this operation.

Once a segment is ready to be sent to the network, the Tx engine uses the checksum 20 stored in the ADE as the checksum for the data portion of the TCP segment.

The Rx engine and the Tx engine use FIFOs 70, 72 for interfacing the engines to the dual port application transfer memory 24. The FIFOs minimize the bus arbitration necessary to read and write data into the dual port application transfer memory 24 from the engines. 25 The control of the FIFOs involves filling and draining the FIFOs in a cycle-steal mode between host system accesses to the memory.

The control unit 44 has an address and data bus connection to the Tx and Rx engines 46, 48. This bus allows the control unit to set and read configuration and status registers within the two engines.

The control unit controls access to the Tx and Rx engines and all memory 24, 50, 52, 30 56, and CAM memory 54 through arbitration, using arbitration unit 74. Host system accesses

and accesses compete for access through the control unit. Any known arbitration method may be used to control these accesses.

Figures 5 and 6, respectively, illustrate in more detail preferred embodiments of the Tx (transmit) engine 46 and Rx (receive) engine 48. Referring to Figure 5, Tx engine 46 may 5 be controlled by a state machine 100, which is used to generate signals which are used to control all the events in the send process. It may be based on a send counter (not shown). This counter is started at initial transmission time, and generates signals which are used to control all the events in the send process. A multiplexer 102 combines Tx data and the outputs of several registers, and provides these to an output register 104.

10 The registers muxed to the output register 104 may be Tx prototype register 106, a Tx application data output register 108, the outputs of checksum registers 110 and 112, an ACK register, and all the individually calculated fields in overlay registers 114 and 116.

15 The send counter and the Tx engine control state machine 100 govern the timeslots for outputting the various fields to the output register. The state machine determines the proper time to calculate the various IP and TCP fields and when to send next segment. Sliding window calculation logic provides the information via register 120 to the Tx engine state machine for next segment transmission.

20 The Tx engine is responsible for sending Ethernet packets containing IP datagrams of TCP segments to the network. There are two primary types of TCP segments. These are user data (ADE) segments, and automatically generated acknowledgment segments for received data. The network accelerator creates packets from scratch, generates the Ethernet header, the IP header, the TCP header, and the TCP data payload.

25 The Tx engine state machine 100, which may be contained in the dual port memory controller 61, asserts the Tx pending state through the Tx engine control state machine 100, making available data contained in the Tx FIFO 70 (Figure 4). The Tx engine loads a prototype register 106 with static portions of the TCP/IP headers from the proto memory 50 of the Tx engine (Figure 3). The logic for the calculation of the dynamic portion of TCP header is contained in a TCP header Tx overlay register 116. The logic for the final checksum calculation for the dynamic portion of the IP header may be contained in an IP header checksum register 110 of Figure 5, and the logic for the post-checksum calculation for

the TCP segment may be contained in a TCP segment post-checksum register 112. The logic which provides sequential accesses to the register contents to the output register 104 may be contained in a transfer register 105. The logic for calculating sequence numbers from the TCP header Tx overlay register by adding the length of the packet data contents to a current sequence number may be contained in an arithmetic logic unit (ALU) 118. The logic to obtain values of a prior received datagram's sequence number and length to generate an acknowledgment number using the ALU may be obtained from the Rx engine TCP header Rx register 103. The final results may be output via the output register 104 to the CRC/MAC unit 42 of Figure 3. The logic for determining whether sending of a datagram was successful and acknowledged may be contained in the engine control state machine 100. The logic for determining if one can send additional datagrams is determined by the engine control state machine 100 and Tx window register 120.

The data used to generate the dynamic calculated portion of the TCP/IP headers reside in the ADE memory, the proto memory, and the memory 24, and the data to generate the static precalculated portion of the TCP/IP headers resides in the Tx engine proto memory 50. The data used to generate the TCP/IP payload resides in the dual port application memory 24.

When the host system asserts a signal indicating that a segment should be sent to the network, the base address of the segment prototype is loaded into the Tx engine proto memory address register, and the base ADE address for the segment is loaded into the Tx engine ADE memory address register. The Tx engine reads the ADE and prototype data out of the ADE memory and the Tx engine proto memory, respectively, then calculates the various fields and inserts the fields into the outgoing network stream. Certain fields of the stream, such as sequence numbers, ACK numbers, ID fields, etc., may be calculated as the stream progresses. Once the headers have been calculated, the TCP payload is output from the dual port application transfer memory. Finally, the CRC/MAC unit 42 appends a CRC 32 value to the Ethernet packet, and completes delivery of the packet to the PHY device 40. In this manner, the network accelerator generates a complete Ethernet packet comprising an IP datagram containing a TCP segment.

Referring to Figure 6, the Rx engine 48 is controlled by an Rx engine control state machine 140, and receives Ethernet packets from the network interface comprising the PHY

device 40 and the CRC/MAC unit 42 via the input register 142. Upon receipt, the state machine sequences data to other elements of the Rx engine. The receive packet is sent to the Rx bypass memory 62, which serves as a buffer and used for any packet that is not a bulk data transfer TCP segment. The Rx engine processes the IP and TCP headers and determines 5 the type of TCP segment. The Vx CAM memory 22 is used by the Rx engine to determine to which session an incoming IP datagram belongs.

The Rx engine, under the control of state machine 140, compares a number of fields of the IP header with expected values stored in a plurality of registers. Certain fields in the TCP/IP header are static and can be compared against static values. Other fields are variable, 10 and define, for example, the length, or checksum or other session-related details. The variable fields are compared against values stored in registers and pre-determined values stored in the ADE memory.

Upon receiving an incoming packet, the header is decoded by a decoder 144 to determine the location of the source and destination addresses and ports contained in the 15 TCP/IP header. The logic for locating the associated prototype packet header and address descripted entry is contained a Vx CAM proto-ADE locator 146. The Rx engine block address decoded entry may be held in an ADE register 156. The Rx engine block prototype entry may be held in proto memory 52, which loads the entry into a prototype register 148. A TCP/IP header matcher 150 which contains logic for comparing of session fields of the 20 packet obtained from the prototype register and variable fields held in a TCP header Rx register 152 and IP header Rx register 154. Logic for validating the checksum for the IP portion of the TCP/IP header matcher 150 may be contained in an IP header checksum unit 162, and the logic which validates the checksum for the TCP segment portion of the TCP/IP header matcher and the data stream from the input register 142 may be contained in a TCP 25 segment header checksum unit 160. Data from valid packets may be passed to the receive data FIFO 72 (Figure 4). Logic for updating TCP header Rx register 152 for transmitted data acknowledgments or buffer window size adjustment may be contained in the arithmetic logic unit (ALU) 164.

The Rx engine control state machine 140 reduces the Rx window register 170 as data 30 is received, and increases it as buffer space becomes available in the dual port application

transfer memory 24 (Figure 3) by the application. Under the control of the state machine, the contents of the Rx window register 170 may also be passed to the ALU 164 to synthesize a window update, which may be passed to the Tx engine via the Rx engine transfer unit 172.

When processing a packet header, if any of the fields of the header do not match 5 expected values, the segment may be routed to the Rx bypass memory 62, and the Rx engine may go into an idle state. The IP source and destination addresses, plus the TCP source and destination ports, may be hashed together to form a value which is used as an address to look up in the content addressable memory the address for the Rx prototype. If the memory returns a non-zero value, it is used as an address to fetch the Rx prototype. If the value is 10 zero, the packet is routed to the bypass buffer.

The value returned by the content addressable memory is used as the base address for the Rx prototype for the segment. The prototype is read and the IP address and the TCP ports are compared against prototype values. If they match, the segment is accepted for further processing, and the ADE base address is read from the prototype memory array. The ADE 15 contains the base sequence number of the memory region. If the sequence number and the segment falls within those in the ADE, it is accepted and the base TCP payload address is read from the ADE.

Data from the segment is read into the dual transport application memory 24 until the segment is completely received, which can be determined by a length counter. Once a 20 segment is received, a CRC 32 signal may be asserted, indicating the packet has been verified and to notify the host system of receipt of data. The Rx engine 48 remains in a pending state until a finished bit is received for the segment. At that time, the system is interrupted and the network accelerator control unit goes into the Rx complete state.

From the foregoing, it may be seen that the network accelerator of the invention 25 affords significant advantages, and may be used in diverse applications. It is also applicable to continuous flow, streamed protocols other than TCP/IP. Some of these applications include high speed links for network backbones, protocol processing for gigabit physical Ethernet layers, data transport between computers within a system, high speed transport for real time high resolution video, increasing the speed of Internet data burst communications,

permitting telephony packets to be transmitted over the Internet, and affording enhanced transaction processing and robotics control feedback.

As will also be appreciated from the foregoing, the implementation of the network accelerator is not limited to FPGAs. It may be implemented in other forms of hardware, and even integrated with microprocessors. The invention may be installed in various components, such as disk drives, graphics cards, video transmission devices, wireless links, TCP/IP hubs, and the like. The substantial increase in speed and corresponding reduction in latency afforded by the invention is a significant advantage.

While the foregoing description has been with reference to particular embodiments of the invention, it will be appreciated by those skilled in the art that changes in these embodiments may be made without departing from the principles and spirit of the invention, the scope of which is defined in the appended claims.

What is claimed is:

1. A method of accelerating a stream-oriented network transport protocol for data packets having a formatted header containing information about the packet, the method comprising analyzing packet traffic on the network to identify classes of predictable protocols which characterize a majority of such packets; implementing programmable hardware logic to process such classes of protocols, said programmable logic being clocked at a rate corresponding to a signaling rate on the network; analyzing the header of a packet to identify one of said classes to which said packet belongs; controlling said programmable logic in accordance with the identified class to process the packets; and processing in software routines instead of said programmable logic packets which do not belong to one of said plurality of classes.

2. The method of Claim 1 wherein said identifying comprises storing in a memory a plurality of predetermined patterns which correspond to said plurality of classes; analyzing the header of a packet to identify a match with a stored pattern; simultaneously with said analyzing, processing the header to determine whether the packet is valid; controlling the programmable logic to process the packet in accordance with the class corresponding to the matched pattern; and processing in said software non-matching and invalid packets.

20

3. The method of Claim 2, wherein said network transport protocol comprises TCP/IP, and wherein said processing comprises simultaneously storing an incoming packet in a dual port memory and in a buffer memory; processing said header during said storing to determine error code values for a TCP segment and for an IP header of the packet, and wherein said identifying comprises comparing such headers and error values with said patterns.

4. A method of accelerating a stream-oriented network transport protocol for data packets having a header, the method being performed by pre-programmed hardware and comprising receiving and storing an incoming packet in a buffer memory; simultaneously

with said receiving and storing, comparing the header of an incoming packet with predetermined patterns stored in a content addressable memory; upon locating a matching pattern, determining a session to which the packet belongs and processing the header to determine whether the packet is valid; and upon finding a matching pattern and determining 5 that the packet is valid, processing the data in the packet in said pre-programmed hardware and delivering to application memory, otherwise forwarding the packet from said buffer memory to a software process for processing as an exception condition of said protocol.

5. The method of Claim 4, wherein said protocol comprises TCP/IP, and wherein
10 said determining steps comprise decoding the header of the incoming packet as the packet is being stored to determine checksum values and TCP and IP header values; and comparing said checksum values and header values to said patterns.

6. The method of Claim 4, wherein said pre-programmed hardware comprises
15 programmable logic which is clocked at a signaling rate of packets on said network.

7. The method of Claim 4, wherein said storing further comprises storing said
packet in a buffer memory and said application memory comprises a dual port, dual bank
memory; and wherein the processing of said packet is controlled by a state machine in said
20 pre-programmed hardware.

8. A network accelerator for a stream-oriented network transport protocol for
data packets having a header comprising means for storing an incoming packet in an
application memory; means for comparing the incoming packet with a plurality of
25 predetermined patterns stored in a content addressable memory to identify a matching pattern;
means for processing the incoming packet simultaneously with said comparing means for
determining whether the packet is valid; means operative upon a matching pattern being
identified and the packet being determined valid for processing said packet in accordance
with the identified pattern; and means operative upon failing to identify a matching pattern or
30 upon determining the packet to be invalid for processing said packet in a software process.

9. The network accelerator of Claim 8 wherein said stream-oriented protocol comprises TCP/IP, and wherein said processing means comprises programmable logic controlled by a state machine.

5

10. The accelerator of Claim 9 wherein said application memory comprises a dual port, dual bank memory, and said accelerator further comprises a control unit including the state machine and control registers for controlling said programmable logic and said memory; said programmable means further comprising means for constructing and storing packets in said application memory; and for supplying packets to said network at the network signaling rate.

1/7

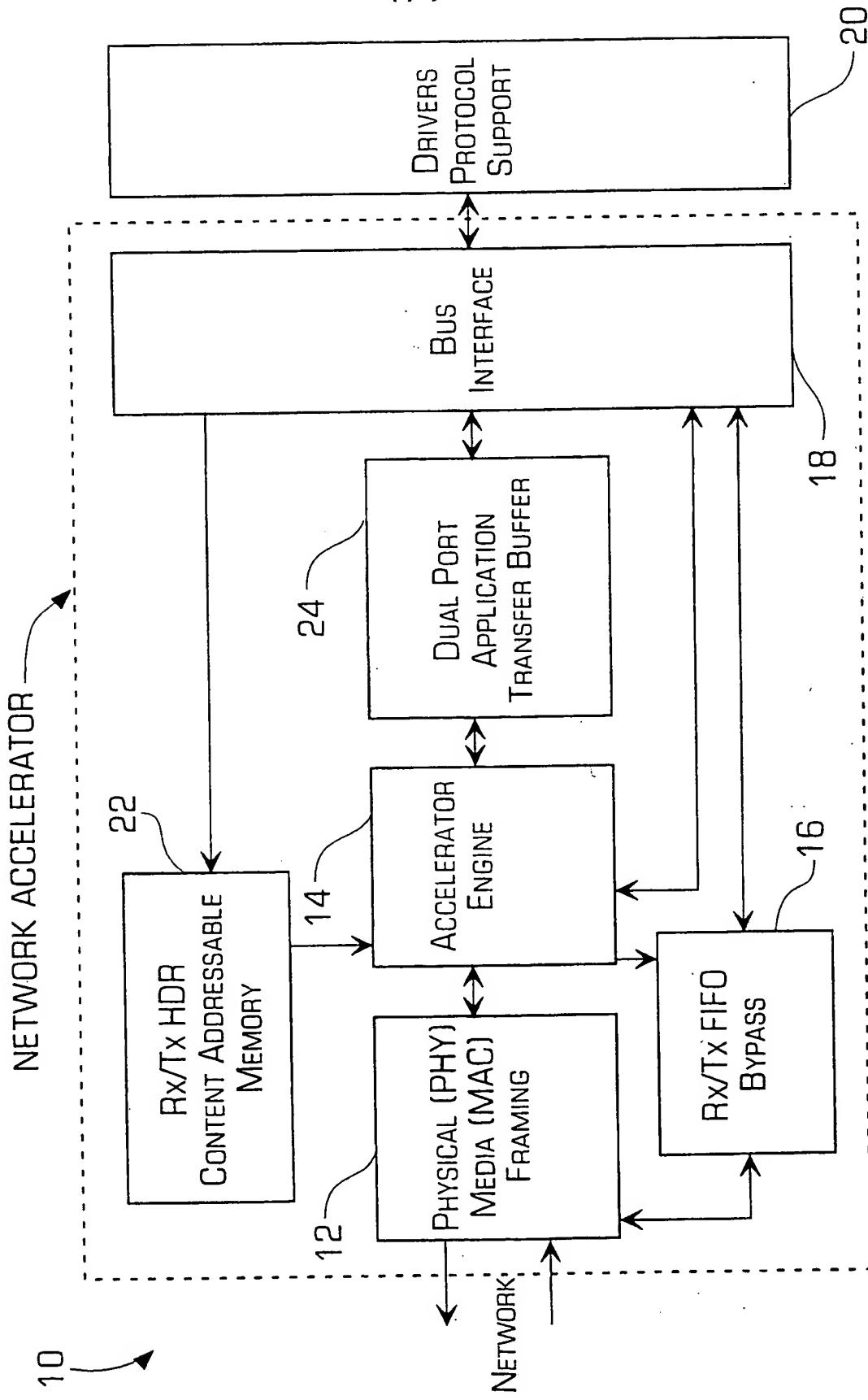


FIGURE 1

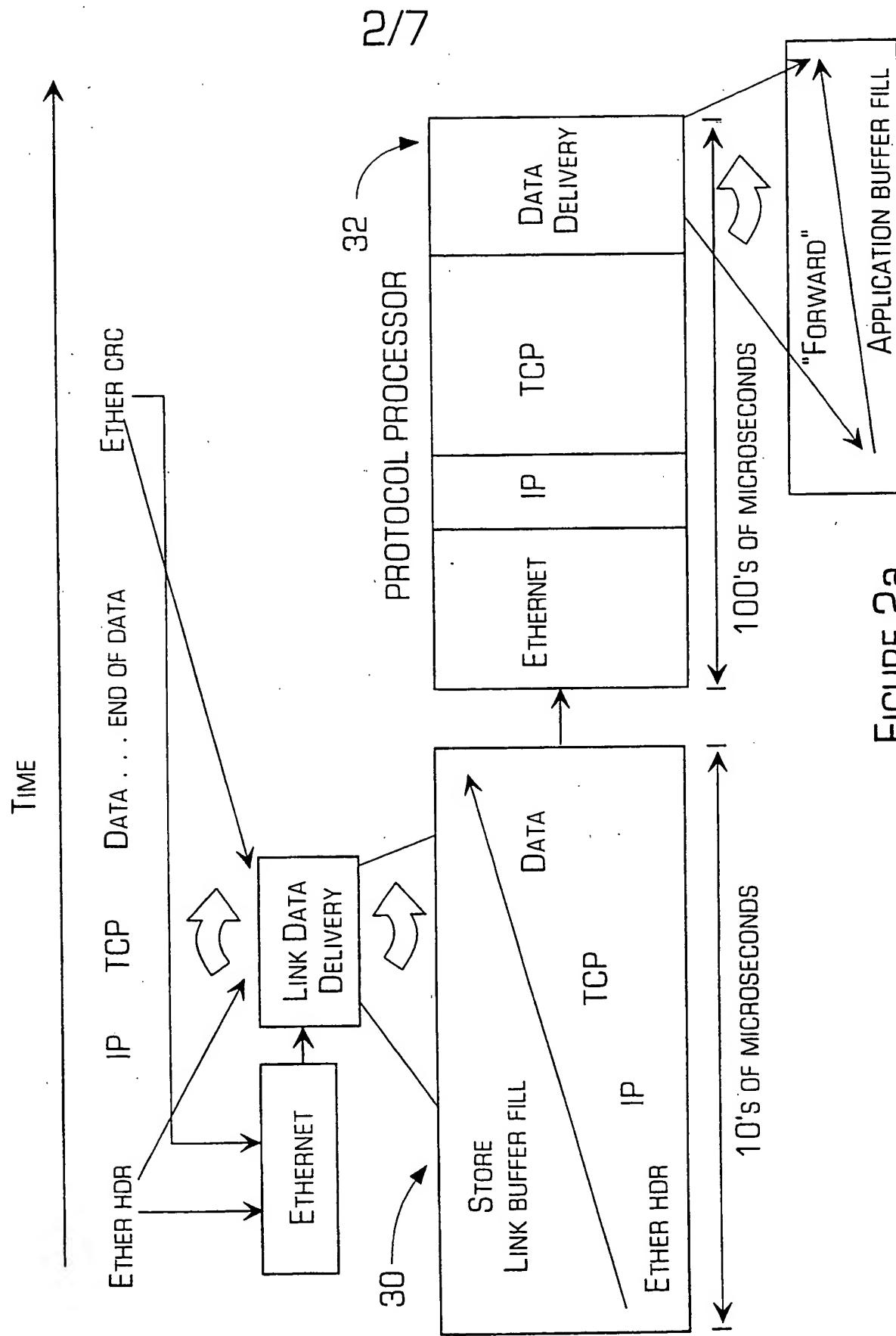


FIGURE 2a

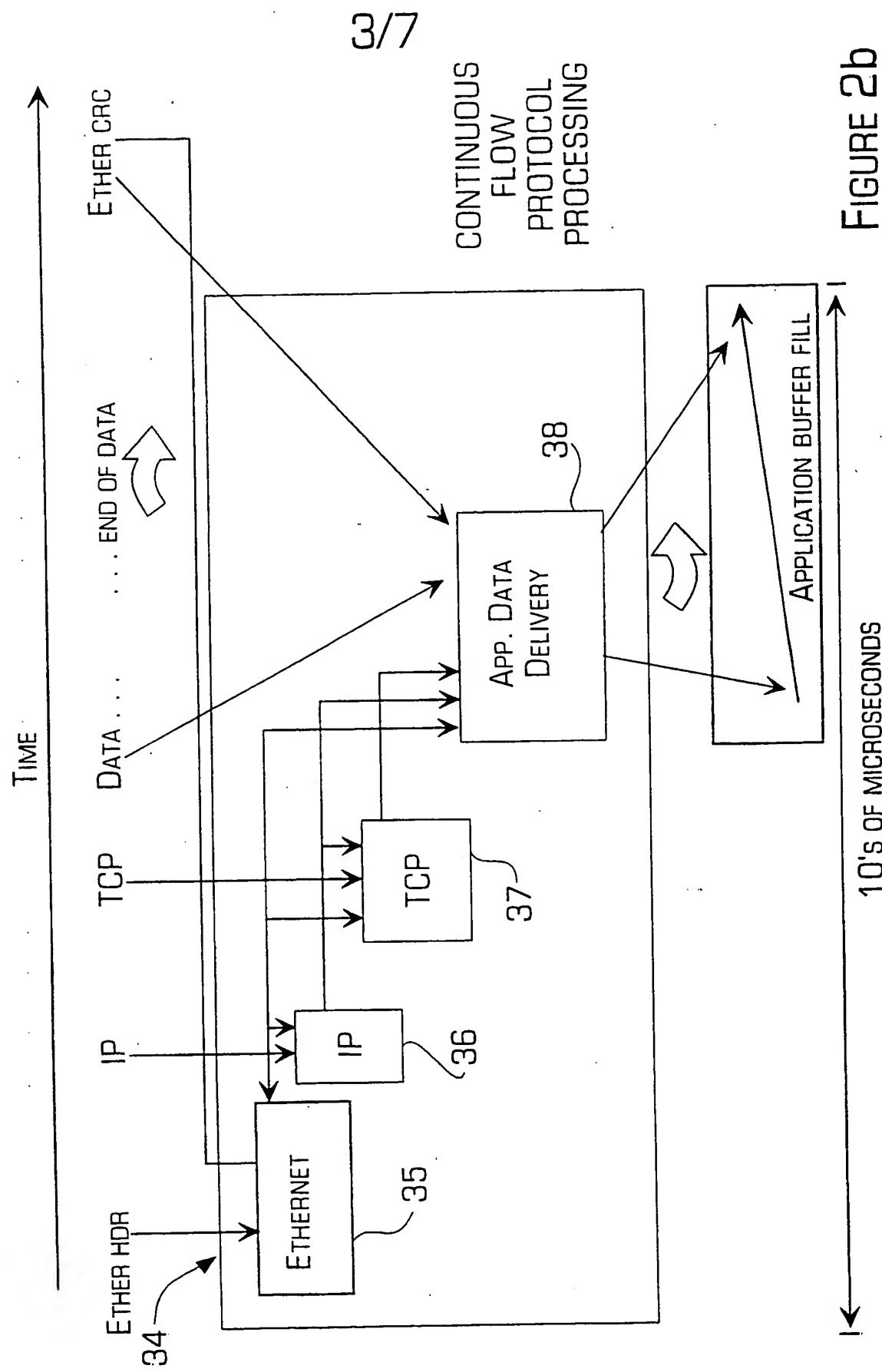


FIGURE 2b

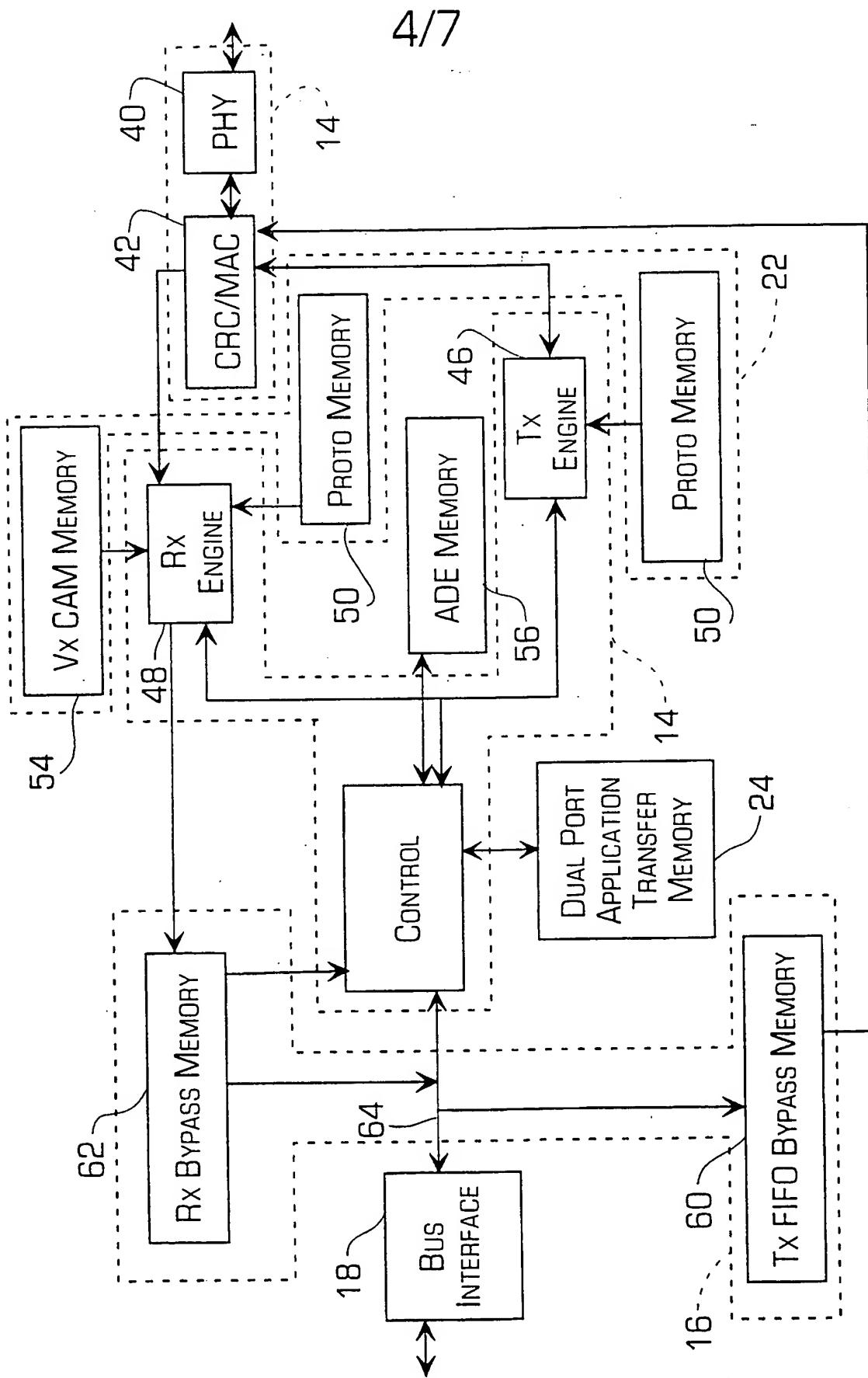


FIGURE 3

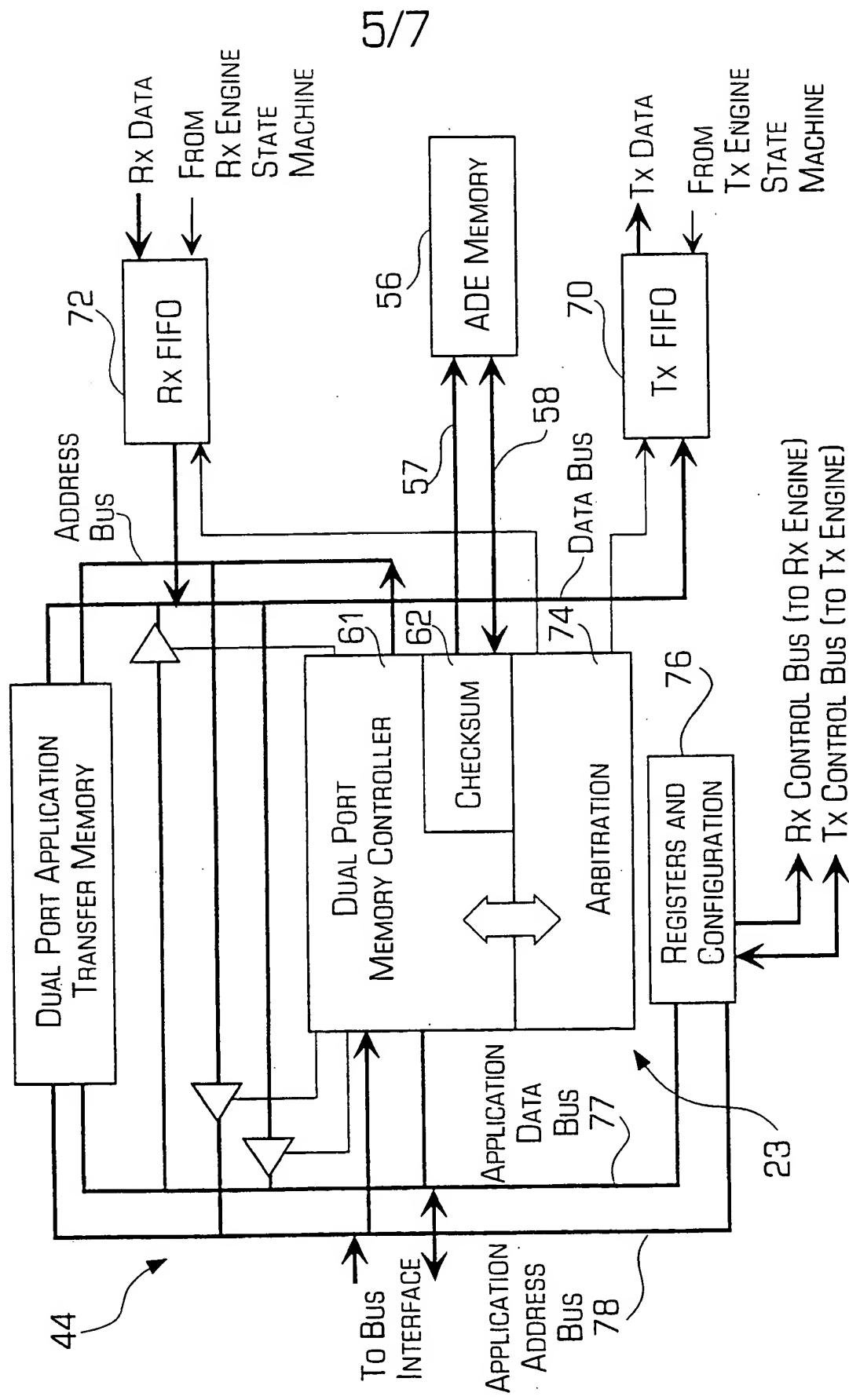


FIGURE 4

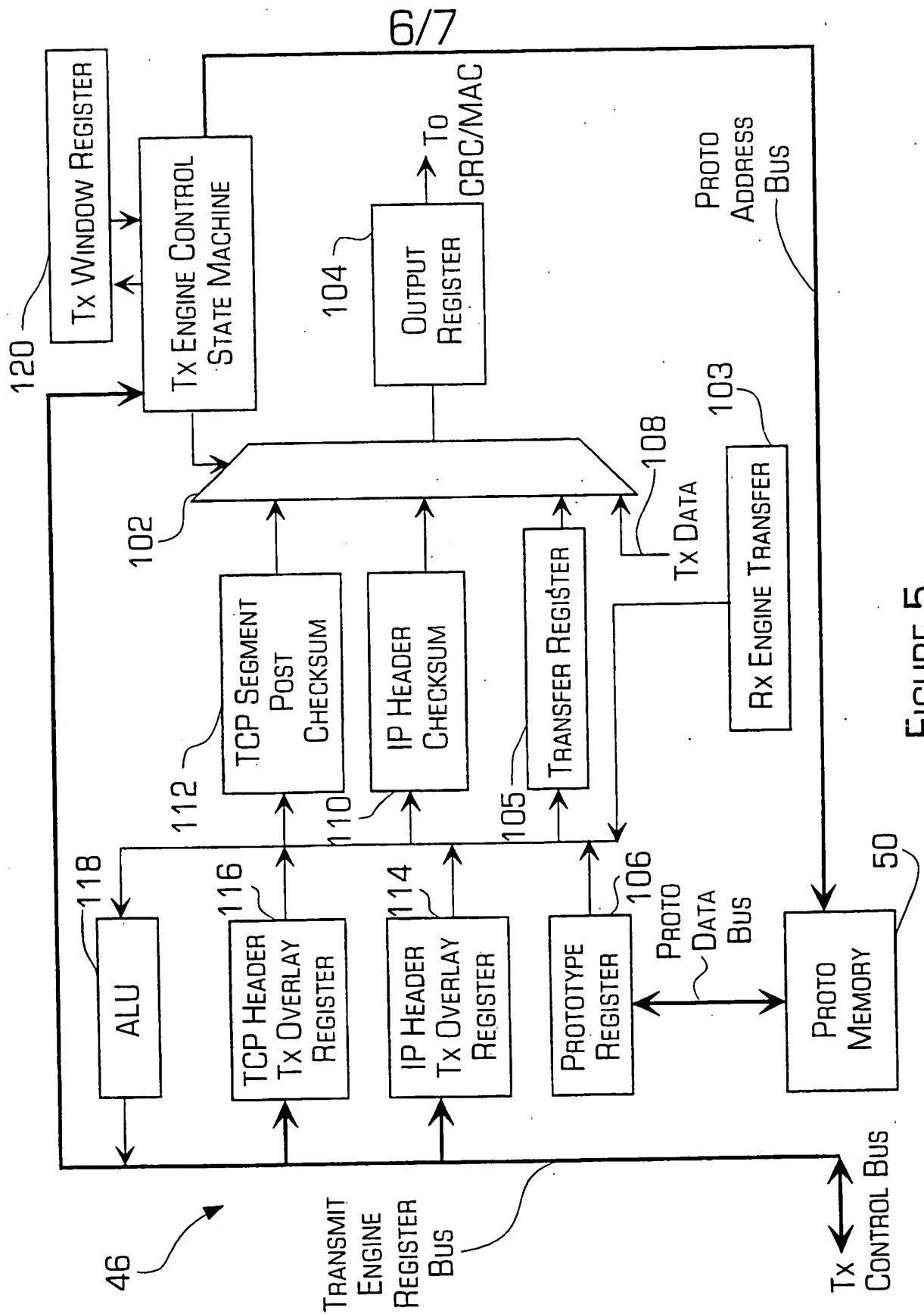


FIGURE 5

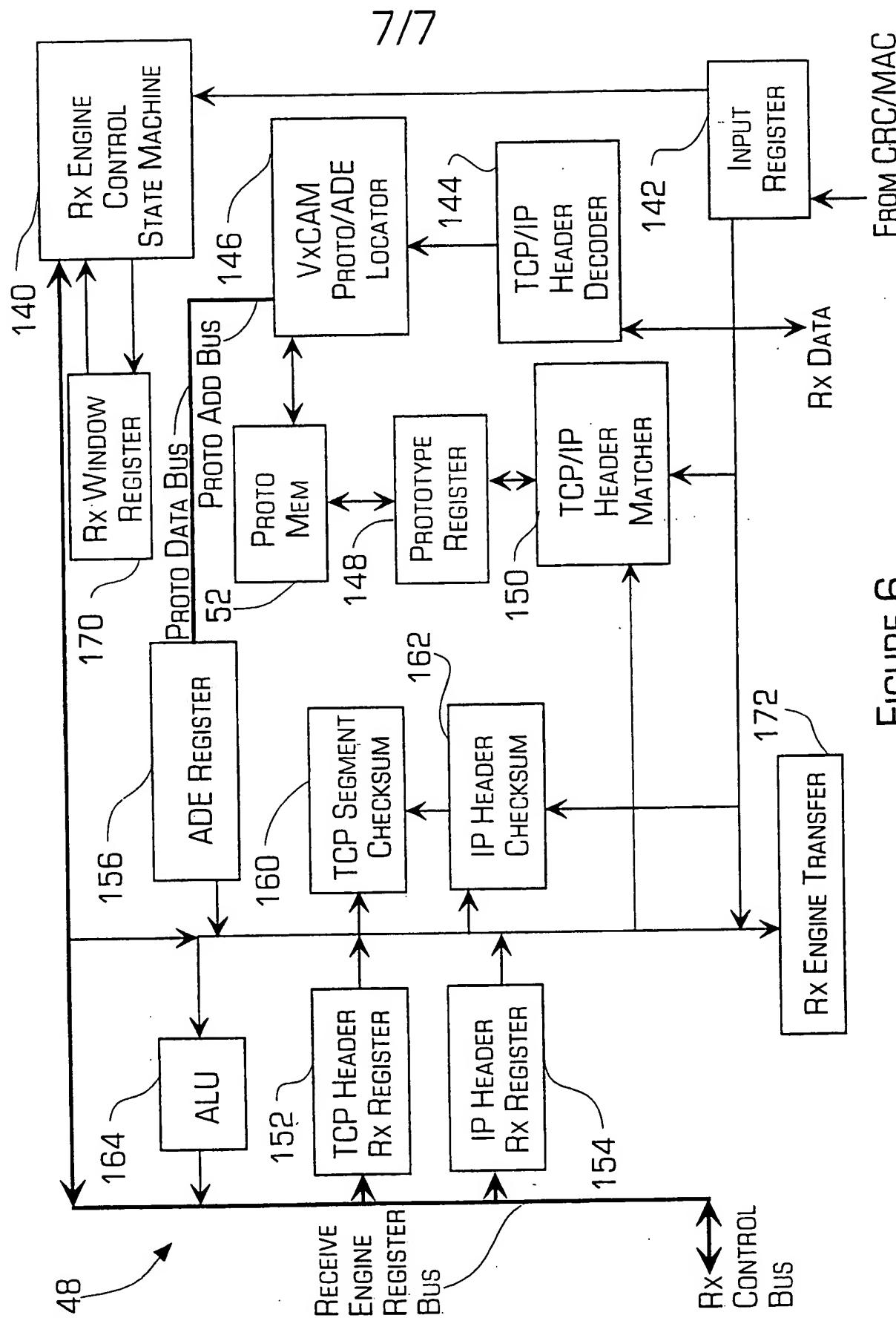


FIGURE 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US98/14729

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : G06F 13/00

US CL : 370/392, 389, 469

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 370/392, 389, 469

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|-----------|---|-----------------------|
| X | US 5,566,170 A (BAKKE ET AL) 15 OCTOBER 1996. ABSTRACT, COL. 4, LINE 60 - COL. 6, LINE 16, COL. 9, LINE 55 - COL. 10, LINE 25, COL. 11, LINE 18 - COL. 12, LINE 49, COL. 13, LINES 24 - 46. | 1-10 |

Further documents are listed in the continuation of Box C.

See patent family annex.

| | | | |
|-------|---|-----|--|
| * "A" | Special categories of cited documents: | "T" | later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention |
| "B" | document defining the general state of the art which is not considered to be of particular relevance | "X" | document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone |
| "L" | earlier document published on or after the international filing date | "Y" | document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art |
| "O" | document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) | "Z" | document member of the same patent family |
| "P" | document referring to an oral disclosure, use, exhibition or other means | | |
| "R" | document published prior to the international filing date but later than the priority date claimed | | |

Date of the actual completion of the international search

06 OCTOBER 1998

Date of mailing of the international search report

09 NOV 1998

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Authorized officer

LE HIEN LUU

